By Christian Collberg and Stephen Kobourov

SELF-PLAGIARISM IN COMPUTER SCIENCE

Should computer scientists reuse their own writing? First, it is critical to determine what qualifies as legitimate reuse and what constitutes self-plagiarism.

We are all too aware of the ravages of misconduct in the academic community. Students submit assignments inherited from their friends, online papermills provide term papers on popular topics, and occasionally researchers are found falsifying data or publishing the work of others as their own.

This article examines a lesser-known but potentially no less bothersome form of scientific misconduct, namely self-plagiarism. Self-plagiarism occurs when authors reuse portions of their previous writings in subsequent research papers. Occasionally, the derived paper is simply a retitled and reformatted version of the original one, but more frequently it is assembled from bits and pieces of previous work.

Illustration by ferruccio sardella

Apart from works by Samuelson [9], Louie [7], Bird [2], and Brogan [3], the legal and ethical implications of self-plagiarism have received little public attention. This is unfortunate since we believe self-plagiarism to be more prevalent than other forms of scientific misconduct and one that can have detrimental effects on our research community:

- It can give the public the idea that research dollars are spent on rehashing old results rather than on original research, simply to further the careers of researchers;
- It can indicate to our colleagues that academic dishonesty is not a big problem. In the worst case this could lead to more serious forms of academic dishonesty becoming more acceptable;
- It rewards authors who break down their results into overlapping least-publishable units over those who publish each result only once; and
- Whenever a self-plagiarized paper is allowed to be published, another, more deserving paper, is not.

Many of us have probably been guilty of some form of self-plagiarism. Maybe we have recycled an introduction or a related work section from one paper to the next. Maybe we have failed to include one of our own related papers in the list of references because it was a bit too related to the topic we were writing on. Maybe we have submitted very similar work to two different communities in order to advertise an important new result.

Here, we present the results of an automated search for self-plagiarism in computer science, recount some personal anecdotes, relate what some of our colleagues think about the issue, and discuss possible ways of addressing the problem.

Three Anecdotes

Our interest in self-plagiarism first arose when A_1 was on the program committee (PC) of C_1. (To preserve anonymity we refer to the authors of the present paper as A_1 and A_2 and the conferences where the incidents described took place as C_1 and C_2.) For one of the papers assigned to A_1, he searched the Web for earlier papers by the same authors that might help to give some background in the area. One paper looked particularly relevant. It had first appeared in a regional conference whose proceedings were later published by a major publisher. On closer scrutiny A_1 found that—except for the title and formatting—the earlier paper was word-for-word identical to the one submitted to C_1. No mention of the original paper was made in the submitted paper.

A_1 alerted the conference chair, who preemptively withdrew the paper from further review. A_1 was also able to submit the shortest referee report of his career: "The following paper should also be referenced [...]."

In a similar incident, A_2 was asked to review a paper for C_2. His review reads as follows:

"Let P_0 and P_1 be the articles [...], published in C_3 and C_4, respectively, and let P_2 be the paper submitted for review. P_0 and P_2 and P_1 and P_2 have intersecting author lists."

"Four pages in P_2 occur identically in P_1. Two figures (one of them containing the major algorithms of the paper) in P_2 also occur in P_1. The bottom of a figure in P_2 (containing a major algorithm) also occurs in P_0. Neither P_0 nor P_1 are referenced in P_2's bibliography."

Such incidents made the first author of this paper think about his own previous papers. Had he ever engaged in self-plagiarism himself? As it turns out, he had once submitted a paper to ACM PLDI'96. It was rejected, and then resubmitted to a regional conference (the 1997 Australasian Computer Science Conference), where it was accepted. He then rewrote the paper, added some major new results, and submitted the new paper to PLDI'97. This time it was accepted. However, he conveniently "forgot" to include a reference to the earlier paper in his submission to PLDI'97. Why? Most likely he was afraid that the previous publication would prevent him from getting the publication in a prestigious international conference he felt he desperately needed.

So, What Happened?

Apparently, not much happened to the authors in these cases. The program chair of C_1 writes:

"I reported the case to [the publisher of the original paper] and got a reply about how they would look into it, but never heard anything more."

"I reported it to the [steering committee of the superstructure to which C_1 belongs] but they did not show much interest."

"I complained to the two authors about their unethical behavior but got no reply. (Not surprising.)"

"I toyed with the thought of complaining to the head of the department where the authors work but didn't do it. The action struck me as being a little vin-

dictive and would achieve little."

The program chair of C_2 writes:

"The paper was rejected and I just sent the review to the authors. Nothing more I'm afraid."

The first author of this paper later saw the errors of his ways, and in the journal version of his paper both conference papers are cited.

Current Policies

Conference call-for-paper announcements and journal submission guidelines usually have a short statement about the use of previously published results. The ACM policy on prior publication and simultaneous submission (www.acm.org/pubs/sim_submissions.html) allows the submission of "papers that appeared previously in refereed publications" provided that: "the paper has been substantially revised (this generally means that at least 25% of the paper is material not previously published; however, this is a somewhat subjective requirement that is left up to each publication to interpret)."

Similarly, the IEEE policy expressly states that plagiarism, self-plagiarism, fabrication, and falsification are "unacceptable" [8]. Both policies give substantial leeway to the journal editor or program chair to decide when a submitted work meets minimum novelty standards. Both policies emphasize novelty of the new result as an important criterion, and ACM puts a number to it: "at least 25% of the paper is material not previously published."

Definitions of Self-Plagiarism

There appears to be little agreement among academics as to what should be regarded as self-plagiarism and what is acceptable republication. We will therefore introduce terminology to allow us to describe the actions performed by authors that might be referred to as self-plagiarism. We will adopt the neutral word *reuse* to refer to texts or ideas that are published multiple times. When appropriate, we have tried to integrate the terminology of others (for example, [6]).

We introduce the following terms:

Textual reuse: Incorporating text/images/or other material from previously published work. (By "published work" we mean articles published in refereed conferences and journals where copyright is assigned to someone other than the author.)

Semantic reuse: Incorporating ideas from previously published work.

Blatant reuse: Incorporating texts or ideas from previously published work in such a way that the two works are virtually indistinguishable.

Selective reuse: Incorporating bits and pieces from previously published work.

Incidental reuse: Incorporating texts or ideas not directly related to the new ideas presented in the paper (such as related work sections, motivating examples, among others).

Reuse by cryptomnesia [4]: Incorporating texts or ideas from previously published work while unaware of the existence of that work.

Opaque reuse: Incorporating texts or ideas from previously published work without acknowledging the existence of that work.

Advocacy reuse: Incorporating texts or ideas from previously published work when writing to a community different from that in which the original work was published.

When these actions pertain to one's own work we talk about textual self-reuse. When it is believed that the actions are ethically or legally questionable we replace reuse by plagiarism, as in blatant semantic opaque self-plagiarism (reusing one's own previously published ideas in a new publication without adequate attribution).

Most would agree that blatant textual self-plagiarism (as exhibited in the submission to C_1) is wrong. Many academics, however, appear to differ in their views of advocacy plagiarism. Certainly, some would argue, it is important to make the public aware of new results relevant to their field, and if a particular result applies to more than one community multiple publications of the same idea are perfectly reasonable. Some make a distinction between horizontal and vertical advocacy reuse. Vertical reuse (republishing a suitably restructured scholarly article in a more "popular" forum) is often deemed acceptable, whereas horizontal reuse (republishing a scholarly article in a similar research forum) is not. The reasoning is that vertical reuse typically does not earn the author academic credit whereas horizontal reuse does.

Incidental plagiarism is also contentious. For example, once we have written the perfect introduction to a problem on a particular topic, can we reuse it for other papers on the same topic or do we need to reword it every time?

And where should the line be drawn between blatant and selective self-plagiarism? The authors of the submission to C_2 reused more than four pages from their previously published work. This is clearly not as bad as trying to republish an entire paper, but neither is it completely aboveboard.

While textual self-plagiarism is easier to detect than semantic self-plagiarism, is it also less ethical? Copy-

right law covers the expression of an idea, not the idea itself, so rewording and republishing a paper (which many would regard as unacceptable) may be perfectly legal.

Searching for Self-Plagiarism

It is difficult to know how common self-plagiarism is. Anecdotally, we hear of colleagues who publish the same result with minor modifications over and over again, and occasionally we come across a paper whose content we feel is too close to previously published work.

We conducted an experiment in which we examined the publications found on computer science Web sites from 50 schools. Our system, SPlaT (Self-Plagiarism Tool) [5], consists of a specialized Web spider and a text-similarity analyzer. SPlaT downloaded the publications of each author from each institution, converted the articles to text, and compared them pair-wise for instances of textual reuse. Works that exceeded a certain threshold were examined manually. (SPlaT can be downloaded from splat.cs.arizona.edu.)

Some highly correlated pairs of papers represented what are generally thought of as acceptable forms of republication: technical reports published in conferences, conference articles recast as journal papers. These were all weeded out manually. However, we found a number of instances of papers with questionable originality. In particular, we ran across cases such as:

- Pairs of conference publications with common introduction and/or related work sections that do not reference each other.
- Pairs of conference publications with over 50% common text that do not reference each other. Note that when measuring textual reuse, we look for sentences and paragraphs that co-occur rather than for words or phrases.
- Pairs of nearly identical conference and journal versions of the same paper, where the journal version does not reference the conference version.

In many ways this is an unsatisfactory study. Not all papers appear on authors' Web sites (in particular, those that have been deliberately self-plagiarized are likely to be absent), not all papers can be successfully parsed, and, finally, it may well be that many attempted cases of self-plagiarism are caught and never make it to publication. However, our study confirms that textual reuse does occur.

What Do Colleagues Think?

To get a feel for how others in the computer science community feel about self-plagiarism, we conducted an informal survey among our colleagues. We sent out 30 questionnaires and received 10 responses. Although no statistical significance can be attributed to this survey, the answers illustrate the range of opinions on this topic.

Our first question wondered if the respondent had encountered cases of self-plagiarism:

(a) "I've seen many cases of papers that served up the same basic ideas in different ways."

(b) "I have encountered submissions where the authors provide the same (word-by-word) introduction, background, and parts of the paper."

(c) "I encountered bad self-plagiarism once. Had I known about the previous publication of the similar work—which I'd guess was 80% identical to submission—I'm sure it would have been killed in committee."[1]

(d) "I'm an associate editor for [...] and we frequently see papers that are more or less complete resubmissions of conference papers. I throw anything out that doesn't make the 30% rule but I believe some of the editors are more 'flexible.'"

(e) "A paper submitted to a good ACM conference by a flourishing research group contained two pages of introduction and two pages of future work identical to that of two other [conference] papers. The paper was rejected on these grounds."

(f) "I got a paper to review, looked up some references and found that the paper in hand was more than half a copy of one of the author's own references. I wrote this in my review. There was no PC meeting, and I was astonished when the paper was accepted."

Our second question asked if the respondents themselves worry about reusing material from previous papers:

(g) "I always rewrite every paper from scratch."

(h) "No, I don't really worry too much about this."

(i) "For sure. I think there's a strong sense that CS papers should be largely self-contained and that inevitably means duplication."

We next sketched a few reuse scenarios and asked whether the respondents thought they would be cause for concern. The first scenario asked about "two conference papers sharing word-for-word introductions

and/or related work sections."

(j) "I think this is very disturbing. As a reviewer, this gives me a very negative impression of the paper and makes me suspicious about the content as well."

(k) "Not a big deal."

(l) "This is something I've done to some extent [...] but the way I deal with it is by thinking that I'll rewrite it if the paper is accepted."

The second scenario asked about "two papers that are essentially the same but sent to different relevant communities (to advertise a single result):"

(m) "This is not acceptable."

(n) "I think this [deserves] public flogging."

(o) "Probably OK, provided the papers are substantially rewritten for the two different communities and reference is made in one paper to the other."

(p) "If I consider the purpose of publication to be the dissemination of results [...] I don't consider this unethical. From the point of view of using publication counting to evaluate performance, such practices probably hurt this system, but since publication counting is idiotic anyway, I don't see this as unethical."

The third scenario asked about "two very related papers (80% or more) and neither of the two cites the other." All respondents agreed this was unacceptable. One respondent wrote:

(q) "I wrote a paper for [a regional conference] then added essentially a results section, and submitted it to [a major conference] where it was accepted. A few years later [a colleague who had been on the conference program committee] asked me about this; I had not cited the [first] paper, nor made the PC aware of this in some other way. I could not remember my motives for this then and I cannot now, except that I am sure that during the submission I was not aware that I was doing anything wrong. Nowadays, with a little more distance, I can just explain it with being young and dumb :-/."

The fourth scenario asked about "a conference paper and a journal version of the same paper that are virtually identical (95% or more)." Most respondents agreed that this was acceptable, or should even be encouraged:

(r) "We need conferences to advertise results quickly and journals to archive those results for longer periods of time."

Our final question asked if the respondents thought self-plagiarism in the computing community is a problem that deserves more attention:

(s) "I think it's a problem, yes, but mainly as a symptom of a deeper problem: the superficiality of the methods used to evaluate academic contribution."

(t) "It is a problem that any program chair or journal editor must be aware of. It would be nice to have some automated way of checking for similar publications when reviewing papers."

(u) "I don't think this is a big problem. Real plagiarism is much worse, as is pre-plagiarism, where someone hears about a new result, either from the inventor or second-hand, and then goes on to reproduce and publish that result himself."

What Can Be Done?

Missing from the ACM and IEEE policy documents is any discussion of what the consequences of ignoring the rules and guidelines might be and whose responsibility it is to prevent plagiarized and self-plagiarized papers from being published. In contrast, most university course syllabi address the definition of plagiarism and who will look for it, as well as its potential consequences.

The program chair of C_1 made a valiant effort to alert the parties involved to the attempted transgression, but received little response. We believe this is indicative of how we as a community view the self-plagiarism issue. We know that it occurs, we deplore that it occurs, but we are not willing to expend the energy to deal with it when it does occur.

It is our belief that self-plagiarism is detrimental to scientific progress. But what should be done to combat it? We pose the following questions to the community:

- Should conferences and journals provide guidelines describing in more detail what practices will be considered self-plagiarism and detailing the process of dealing with such practices?
- Should the burden of detecting and dealing with plagiarism and self-plagiarism be on professional organizations and publishers?
- Should paper reviewers become plagiarism police?[2]
- What should the consequences be when we find a

[1]Private correspondence indicated that anecdotes (c) and (q) refer to the same incident.

[2]This is the practice in some fields. For example, the *Journal of Advanced Nursing* asks reviewers: "Consider here whether 'salami slicing' of publications, or plagiarism (including self-plagiarism) are possibilities. Has this paper (or parts of it) been published before? Can you identify any potential copyright problems?"

paper we feel has been self-plagiarized? The chair of C_2 simply rejected the paper although several pages had already appeared in print. The chair of C_1, on the other hand, felt that the unethical nature of the incident warranted further actions, but stopped short of reporting to the author's department chair.

It is our belief that we should hold ourselves to the same high standards as we do our students. Many professors use tools such as moss [1] or Glatt Plagiarism Services (www.plagiarism.com) to detect plagiarism among students. Similar tools would be useful to detect self-plagiarism among academics. We are currently modifying SPlaT to act as a reviewer's workbench. The program will compare a paper under review to a record of the author's previously published articles extracted from their Web site and online article repositories. C

References

1. Aiken, A. A system for detecting software plagiarism; www.cs.berkeley.edu/~aiken/moss.html.
2. Bird, S.J. Self-plagiarism and dual and redundant publications: What is the problem? Commentary on "Seven ways to plagiarize: Handling real allegations of research misconduct." *Sci. Eng. Ethics 8*, 4 (2002).
3. Brogan, M. Recycling ideas. *Coll. Res. Libr. 52*, 5 (Sept. 1992), 453–464.
4. Carpenter, S. Plagiarism or memory glitch? Inadvertent plagiarism complicates efforts to end cheating. *Monitor Psychol. 33*, 2 (Feb. 2002).
5. Collberg, C.; Kobourov, S.; Louie, J.; and Slattery, T. SPlaT: A system for self-plagiarism detection. In *Proceedings of the IADIS International Conference on WWW/Internet*, 2003, 508–514.
6. Evans, J. The new plagiarism in higher education: From selection to reflection. *Interactions 4,* 2 (2000); www.warwick.ac.uk/ETS/interactions/vol4no2/index.htm.
7. Loui, M.C. Seven ways to plagiarize: Handling real allegations of research misconduct. *Sci. Eng. Ethics 8,* 4 (2002).
8. Publications standards policy and principles for authors, referees, and editors; www.ieee.org/portal/index.jsp?pageID=corp_level1&path=about/whatis/policies&file=p6-4.xml&xsl=generic.xsl.
9. Samuelson. P. Self-plagiarism or fair use. *Commun. ACM 37,* 8 (Aug. 1994), 21–25.

Christian Collberg (collberg@cs.arizona.edu) is an assistant professor in the Department of Computer Science at the University of Arizona, Tucson.

Stephen Kobourov (kobourov@cs.arizona.edu) is an assistant professor in the Department of Computer Science at the University of Arizona, Tucson.